DOCUMENT RESUME

ED 389 748 TM 024 381

AUTHOR Meijer, Rob R.
TITLE Nonparametric and Group-Based Person-Fit Statistics: A Validity Study and an Empirical Example. Research Report 94-12.
INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE Nov 94
NOTE 38p.
AVAILABLE FROM Bibliotheek, Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; Classification; College Students; *Cutting Scores; Foreign Countries; *Group Membership; Higher Education; Identification; Knowledge Level; *Nonparametric Statistics; Research Methodology; Simulation; *Test Items; *Validity
IDENTIFIERS Aberrance; *Person Fit Measures; *Power (Statistics)

ABSTRACT

In person-fit analysis, the object is to investigate whether an item score pattern is improbable given the item score patterns of the other persons in the group or given what is expected on the basis of a test model. In this study, several existing group-based statistics to detect such improbable score patterns were investigated, along with the cut scores that have been proposed in the literature to classify an item score pattern as aberrant. Through a simulation study and an empirical study, the power of three person-fit statistics was compared, and the practical use of various cut scores was investigated. The empirical study involved 437 Dutch sophomores studying psychology and pedagogics taking an examination on test theory. It was also demonstrated that person-fit statistics can be used to detect persons with a deficiency of knowledge on an achievement test. While one of the statistics was less appropriate in the simulation, the power of the three approaches was approximately the same in the empirical example. (Contains 9 tables and 31 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 389 748

Nonparametric and Group-Based Person-Fit Statistics: A Validity Study and an Empirical Example

Research Report 94-12

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. NELISSEN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

Rob R. Meijer

BEST COPY AVAILABLE

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

TM024351

Nonparametric and Group-Based Person-Fit Statistics: a Validity Study and an Empirical Example

Rob R. Meijer

Nonparametric and group-based person-fit statistics: A validity study and an empirical example, Rob R. Meijer - Enschede: University of Twente, Faculty of Educational Science and Technology, November, 1994. - 32 pages.

Abstract

In person-fit analysis it is investigated whether an item score pattern is improbable given the item score patterns of the other persons in the group or given what is expected on the basis of a test model. In this study several existing group-based statistics are discussed to detect such improbable item score patterns, along with the cut scores that have been proposed in the literature to classify an item score pattern as aberrant. By means of a simulation study and an empirical study the power of these statistics is compared and the practical use of various cut scores is investigated. It is furthermore demonstrated that person-fit statistics can be used to detect persons with a deficiency of knowledge on an achievement test.

Index terms: aberrance detection, appropriateness measurement, group-based person-fit statistics, nonparametric item response theory, person-fit research, person-fit statistics.

**Nonparametric and Group-Based Person-Fit Statistics:
a Validity Study and an Empirical Example**

To assess an examinee's overall ability level on an aptitude or achievement test, usually the (weighted) number-correct score on the test is used. To obtain more detailed information about a person's test performance, the patterns of individual item scores may be studied. Information concerning these patterns may be useful for diagnostic purposes. For example, a person may answer the most difficult items correctly, but the easiest items incorrectly. Such a pattern may be the result of cheating. Other examples of information that may be obtained by studying score patterns deal with information about guessing, plodding, or lack of abilities. (For an overview of several kinds of aberrant response behavior see Hulin, Drasgow, & Parsons, 1983, chap. 4.) Research which is concerned with the investigation of individual item score patterns is known as appropriateness measurement research or person-fit research. Many methods for the detection of aberrant item score patterns have been proposed (e.g., van der Flier, 1982; Levine & Drasgow, 1982; Levine & Rubin, 1979; Molenaar & Hoijtink, 1990; Tatsuoka, 1984; Tatsuoka & Tatsuoka, 1983; Trabin & Weiss, 1983; Wright & Stone, 1979, pp. 165-180). Most of these methods are developed in the context of item response theory (IRT).

In IRT a number of models have been proposed in which probabilities of item responses are explained by the characteristics of a person (the latent ability) and the characteristics of the items (e.g., the item difficulty) (e.g., Hambleton & Swaminathan, 1985). In IRT, parametric and nonparametric models can be

distinguished. In parametric IRT models (Birnbaum, 1968; Rasch, 1960), the probability of a correct item response is a parametrically defined function of the item and person parameters, and for the purpose of parameter estimation sometimes the distribution of the person parameter is parametrically specified as well. In nonparametric IRT models (e.g., Holland, 1981; Mokken, 1971; Stout, 1990) the probability of an item response is not parametrically defined. Besides, the distribution of the person parameter is not restricted to a particular distribution form, but the exact form is left free. Several differences exist with respect to nonparametric and parametric IRT models. For example, nonparametric models are in general less restrictive than parametric models. As a consequence using nonparametric models more items may fit the model than using parametric models. However, nonparametric models allow only measurement on an ordinal scale, whereas parametric models allow measurement on a metric scale. Furthermore, using parametric models the evaluation of measurement precision can be obtained as a function of the latent attribute by means of the information function. Nonparametric models do not allow numerical estimates of person and item parameters that are needed to estimate the information function. Therefore, accuracy of measurement is only determined for the whole population. For fit procedures of parametric and nonparametric IRT models to empirical data see, for example, Meijer, Sijtsma, and Smid (1990).

In a population of individuals item score patterns that are in concordance with the IRT model (normal patterns) can be explained by a person's ability, attitude, or trait level, and the characteristics of the items. Consequently, studying these patterns will not add information to the number-correct score. Therefore, in person-fit research the quest is always for the unusual, aberrant, deviant, or inappropriate patterns.

In person-fit research three groups of statistics can be distinguished (Meijer, 1994; Meijer & Sijtsma, in press). In the first group, an observed item score pattern is evaluated against the predicted pattern under certain restrictions derived from a specific parametric IRT model such as the 3-parameter logistic model or the Rasch model. Examples of this approach can be found in Levine and Drasgow (1982), Drasgow, Levine, and Williams (1985), Drasgow, Levine, and McLaughlin (1987) and Molenaar and Hoijtink (1990). In the second group an item score pattern is evaluated given that a nonparametric IRT model can be applied to the data, such as the Mokken models (Meijer et al., 1990; Mokken & Lewis, 1982). Examples of this approach can be found in van der Flier (1980, 1982) and Meijer, Molenaar, and Sijtsma (in press). Finally, there is a group of statistics that can be used to evaluate an item score pattern without model assumptions (Harnisch, 1983; Harnisch & Linn, 1981).

Recently, much theoretical research has been done in the context of specific parametric IRT models. For example, much research has been conducted in the context of the 3-parameter logistic model on the power of a log-likelihood statistic (Drasgow et al., 1985, 1987; Reise & Due, 1991). For this statistic a sampling distribution has been derived through which it can be decided to classify an item score pattern as either normal or aberrant. Molenaar and Hoijtink (1990) derived several sampling distributions for a log-likelihood statistic in the context of the Rasch model.

If a nonparametric IRT model is used to describe the data, or if no IRT model at all is used, person-fit statistics from the parametric approach can not be applied and statistics from the second and third group should be used.

A practitioner who applies these statistics to detect aberrant persons may want to know when he or she can label a pattern as normal or aberrant. There-

fore, in several studies (Harnisch & Linn, 1981; Miller, 1986) cut scores are reported: persons with statistic values under or above the cut score (depending on whether low or high values indicate aberrant behavior) are then labelled as aberrant. However, a major shortcoming of these studies is that these cut scores are often based on a limited number of empirical data sets. For example, Harnisch and Linn (1981) proposed a cut score for their modified caution index that was based on only two empirical data sets. Little is known about the numerical values of these statistics across varying item and test characteristics. Ideally, a sampling distribution should be derived for these statistics. However, because no parametric IRT model is assumed to underlie the item responses and hence the probability of an item score pattern can not be derived from a model this seems difficult to realize. Another problem is that on the basis of the literature it is difficult to decide which statistic should be preferred to detect aberrant persons. A majority of the literature on parametric person-fit statistics has concentrated on this question by investigating the power of several statistics to detect aberrant item score patterns (e.g., Drasgow et al., 1987). However, using nonparametric and group-based person-fit indices there are almost no studies available.

The purpose of this study was twofold. First, it was investigated by means of simulated data whether the classification rules suggested for some nonparametric and group-based statistics can be applied meaningfully. Second, the power of these statistics to detect aberrant item score patterns was compared by means of both simulated and empirical data.

Nonparametric and Group-Based Statistics

Several nonparametric and group-based statistics have been proposed. Harnisch and Linn (1981) compared eight group-based statistics with respect to their correlation with the number-correct score. They concluded that several of these statistics are less suited to detect aberrant item score patterns because they had a high negative or positive correlation with the number-correct score on the test. As a result, in particular persons with a low or a high number-correct score will be classified as aberrant by these statistics. Since persons across all score levels may generate deviant item score patterns, a high correlation with the number-correct score is undesirable. From the study by Harnisch and Linn (1981) and from a study by Miller (1986) two statistics for which a cut score was proposed were selected. A third statistic was chosen from the work by van der Flier (1982).

The Caution Index. Sato (1975; cited in Harnisch and Linn, 1981) proposed the caution index (C). This index is defined as the complement of the ratio of two covariance terms. Assume that k items are ordered according to their proportion correct score, π_g (g = 1, ..., k), such that $\pi_1 \geq \pi_2 \geq ... \geq \pi_k$. Let $\underline{X}$ be a vector containing the observed binary item scores of an individual person, and let $\underline{X}^*$ be a vector of a person with a number-correct score X = r, with 1's in the first r positions and 0's in the last k - r positions. This vector is called a Guttman vector because it fits the Guttman (1950) model. Furthermore, let $\underline{n}$ be the vector containing the number of correct answers on each of the k items in a test across persons (item-total vector). If $\sigma(...)$ denotes the covariance between

the elements of two vectors, then C is given by

$$C = 1 - \frac{\sigma(\underline{X},\underline{n})}{\sigma(\underline{X}^*,\underline{n})} . \qquad (1)$$

C equals 0 if an individual's item score vector is a Guttman vector; C equals 1 if the covariance between the item score vector and the item-total vector equals 0; if the covariance between the item score vector and the item-total vector is negative, C exceeds 1. Thus, relatively large values of C indicate that the response behavior is atypical compared with the response behavior of the other persons in the group. For C a cut value of .5 was suggested (Sato, 1975; Miller, 1986); item score patterns for which C>.5 should be classified as aberrant.

The Modified Caution Index. Because C has no fixed upper bound, the interpretation of the values of C may be problematic (Harnisch and Linn, 1981). Therefore, Harnisch and Linn (1981) proposed the modified caution index (C^*). The values of C^* range from 0 to 1. Let $\underline{X}'$ denote the item score vector with 0's in the first k - r positions and 1's in the last r positions. Given X = r it is the vector with the maximum number of Guttman errors, and therefore is called a reversed Guttman vector. Then C^* can be written as

$$C^* = \frac{\sigma(\underline{X}^*,\underline{n}) - \sigma(\underline{X},\underline{n})}{\sigma(\underline{X}^*,\underline{n}) - \sigma(\underline{X}',\underline{n})} . \qquad (2)$$

C^* equals 0 if the covariance between an individual item score vector and the item-total vector equals the covariance of a perfect Guttman vector (same number-correct score) with this vector; thus if $\underline{X} = \underline{X}^*$. C^* equals 1 if the covariance of the item score pattern with the item-total vector equals the covariance

of a reversed Guttman vector (same number-correct score) with the item-total vector; thus if $\underline{X} = \underline{X}'$.

Harnisch and Linn (1981) preferred C^* to the other person-fit statistics in their study because C^* was least confounded with the number-correct score in two empirical data sets. For C^* a cut value of .3 was used: item score patterns with $C^*>.3$ were classified as aberrant.

The U3 Statistic. In the context of a nonparametric IRT model (Mokken & Lewis, 1982) that assumes that persons can be ordered according to their ability level, van der Flier (1980, 1982) developed the person-fit statistic U3. Let P, in general, denote a probability, and let $P(\underline{X})$ denote the probability of an item score pattern $\underline{X}$ conditional on the number-correct score. U3 is defined as

$$U3 = \frac{\ln P(\underline{X}^*) - \ln P(\underline{X})}{\ln P(\underline{X}^*) - \ln P(\underline{X}^I)} . \qquad (3)$$

U3 ranges from 0 to 1, where 0 indicates that the observed item score pattern is a Guttman vector and 1 indicates that the observed pattern is a reversed Guttman vector. Increasing values indicate that patterns are farther removed from perfect Guttman patterns. Van der Flier (1980, 1982) derived a sampling distribution for a standardized version of U3, denoted ZU3; ZU3 is approximately standard normally distributed given an invariant ordering of persons along the scale.

By means of a simulation study (van der Flier, 1982), characteristics of ZU3 were investigated. It was concluded, that for sets of 17 and 29 items with proportion correct score values that are either uniformly or normally distributed, the ZU3 distributions within different score groups could be combined into one common distribution. For a study of the influence of several item, test, and person characteristics on the power of U3 see Meijer et al. (in press).

Method

First, data matrices consisting of 3000 normal persons and 17 or 33 items were generated using 3-parameter logistic item response functions (IRFs) and a standard normal distribution for the latent attribute, Θ. Item discrimination (a) was randomly drawn from a ~ U[.5,1.5] and item difficulties (b) were equidistant with a distance equal to .25 for the 17-item test and .125 for the 33-item test. For each test the median difficulty equalled 0. The guessing parameter was randomly drawn from a uniform distribution c ~ U[0,.4]. Both for the 17 and 33-item test for each person C, C^*, and ZU3 were calculated and persons were ordered from the lowest (V_1) to the highest (V_{3000}) value of each statistic. Let α be the probability of misclassifying a normal person as aberrant under the 3-parameter logistic model. Then the 90th, 95th, and 99th percentile values of each statistic were approximated by the value of $V_{3000(1-\alpha)+1}$, where α = .10, .05, and .01, respectively. These values were used as cut scores in the aberrant samples described below. Note that although our theoretical framework is nonparametric and group-based indices, parametrically defined IRFs and parameter distributions are necessary to simulate 0's and 1's.

Second, data matrices consisting of 2000 aberrant persons were simulated. Two kinds of aberrant behavior were simulated: guessing and cheating. Guessing simulees were assumed to answer the items by blindly guessing for the correct answer on each of the k items in the test with a probability of .25 (which corresponds with the probability of obtaining the correct answer by blindly guessing in a multiple-choice test with four-choice items). Cheating simulees

answered most items on their own (item scores simulated by means of the 3-parameter logistic model) except for the three most difficult items from the 17-item test and the six most difficult items from the 33-item test. The answers on these items were assumed to be copied from more able neighbors taking the same test. It was assumed that this cheating always resulted in correct answers and 1's were thus substituted for these item scores for each cheater.

With respect to the detection rate within a data set, the percentages of a priori defined aberrant simulees successfully identified by means of C, C^*, and ZU3 were determined using the cut scores obtained in the normal samples. The main criteria for evaluating the three statistics were the proportions of aberrant simulees (valid aberrants) that were correctly classified when various proportions of normal simulees were misclassified as aberrant (false aberrants). Furthermore, the correlation of the statistics with the number-correct score was determined.

Results

The Tables 1 (k = 17) and 2 (k = 33) show the critical values for C, C^*, and ZU3 at a 1%, 5%, and 10% error rate, respectively.

Insert the Tables 1 and 2 about here

At a 1% error rate the differences between the critical values for C for k = 17 and k = 33 are almost .2, whereas the critical values for C^* and ZU3 are almost the same for the two test lengths. At a 5% and 10% error rate the differences for

all three statistics are negligible. It is interesting that the critical value used by Harnisch and Linn (1981) for C^* (C^* = .3) allows a 10% error rate in a normal sample generated under the 3-parameter logistic model and $\Theta \sim N(0,1)$. Furthermore, it was found that the critical value of .5 for C (Sato, 1975; Miller, 1986) corresponds to the 83th percentile value for the 17-item test and the 85th percentile value for the 33-item test.

The Tables 3 and 4 show the percentages of valid aberrants for C, C^*, and ZU3 for k = 17 and guessing and cheating simulees.

Insert the Tables 3 and 4 about here

Note that at a 1% error rate the detection rates for all three statistics are quite moderate both for guessing and cheating simulees (detection rates between 38% and 44%). At a 10% error rate the detection rates vary between 77% and 91%. Increasing the test length to 33 items (Tables 5 and 6) the percentages of valid aberrants incline sharply, both for the guessing and cheating simulees. This is in agreement with the results found in a study by Reise and Due (1991). They found that the power of a log-likelihood statistic increases as the test length increases. At a 1% error rate the detection

---

Insert the Tables 5 and 6 about here

---

rates are approximately 70% for the guessing simulees and 80% for the cheating simulees. At a 10% error rate these rates are 90% for guessing simulees and almost one 100% for cheating simulees.

If we compare the power of the three statistics there seems to be a trend that at a fixed error rate the power of C is almost always less than the power of C^* and ZU3; C^* and ZU3 are in most cells about equally effective in detecting aberrant simulees.

The correlation with the number-correct score for the three statistics in both the normal and aberrant samples varied between -.03 and -.30. With respect to C^* and C this was in agreement with most of the correlations found by Harnisch and Linn (1981) for these statistics. They found for C^* a correlation of -.02 with the number-correct score on a math test and -.21 with the number correct score on a reading test; whereas for C a correlations of -.17 and -.42 were found. However, unlike the results by Harnisch and Linn (1981), in this study C^* had not always the lowest correlation of the three statistics; all correlations fluctuated around the -.20 with the number-correct score.

The results obtained so far indicate that the power of C^* and ZU3 is somewhat higher than the power of C. Besides, the critical values at a 1% error rate for C^* and ZU3 seem to be more stable across different test lengths. However, the number of normal and aberrant persons and the kind of aberrant behavior were known a priori. In practice both the ratio of normal/aberrant persons and the kind of aberrant behavior are unknown. Therefore, it may be interesting to investigate the use of these statistics by means of an empirical example.

An Empirical Example

It should be realized that not every item score pattern that is identified as aberrant by a person-fit statistic can be interpreted meaningfully. In most cases it will be rather difficult to recognize, for example, a guessing or a cheating pattern. In realistic test situations, aberrant response behavior may be difficult to recognize because (1) due to the probabilistic nature of the process underlying the item responses, item score patterns may not convincingly reflect the underlying aberrant behavior, and/or (2) aberrant behavior may only play a part in a small number of items from the test. In addition, even if a pattern is statistically identified as aberrant, the researcher can not always be sure of the kind of aberrance underlying test performance because different forms of aberrant behavior may result in the same kind of item score pattern.

In general there are three ways to detect persons with a specific kind of aberrant behavior. A first option is to use specialized scales such as scales that are sensitive to lying or faking. A second option is to use specialized statistics that are sensitive to a particular kind of aberrant behavior, for instance, one of the statistics that are discussed by Frary (1993) for detecting answer-copying. A third alternative is to use person-fit statistics and (1) to construct the test in such a way that the values of the statistics can easily be interpreted and (2) to relate the statistics to other experience variables. This third approach was followed in the next example.

Data

To investigate the power of C, C^*, and ZU3 to detect aberrant item score patterns, empirical data of 437 Dutch sophomore students in Psychology and Pedagogics were used on an examination on test theory. The examination consisted of two kinds of items. One group of items was concerned with different kinds of arithmetic skills such as the calculation of raw and standard scores, the calculation of several kinds of reliability coefficients (such as Cronbach's alpha, KR-20), and the calculation of the Spearman-Brown formula. The other group of items were knowledge questions concerning test theoretical subjects.

The exam consisted of 40 items. Item analysis revealed that all item-test correlations were positive and that most item-test correlations were between .1 and .4. Cronbach's alpha equalled .81. Further item analysis showed that three items seriously violated the assumptions of nondecreasing IRFs. These three items were removed from the test. The final test consisted thus of 37 items of which 17 arithmetic items and 20 knowledge items.

From the results on earlier examinations we knew that for a majority of examinees the arithmetic items were easier than the knowledge questions (in general a higher proportion correct-score on the arithmetic items than on the knowledge items). This was also the case for this examination. From the literature (e.g., van der Flier, 1982) it is known that person-fit statistics may only be effective to separate normal from aberrant behaving persons if there is a minority of persons that behave differently from a majority of persons. Because there was a majority of examinees for whom the arithmetic items were easier than the knowledge items, it seemed interesting to investigate if there exist a group of persons for whom the opposite apply.

Method

C, C^*, and ZU3 were calculated to investigate if the results found in the simulation study also apply to this empirical example. For C, C^*, and ZU3 the cut scores given in Table 2 were used. These cut scores were determined for tests consisting of 33 items that were in concordance with the 3-parameter logistic model. These test characteristics were close to the test characteristics of the empirical data.

Results

Table 7 shows the percentages of examinees that were classified as aberrant using the three cut scores for each statistic. For each cut score these percentages are somewhat higher than on the basis of a normal sample could be expected. This is not surprising because this sample may contain aberrant examinees. Note that all three statistics classified approximately the same number of examinees as aberrant. Table 8 shows the mean proportion correct score on the knowledge items and on the arithmetic items in the group with normal examinees and in the group with aberrant examinees. Only the cut scores belonging to the 90th percentile values were used so that the number of aberrant persons was high enough to obtain mean proportions that could be interpreted meaningfully. It can be seen that in the group of normal examinees the knowledge items were more difficult than the arithmetic items. In the group of aberrant examinees the opposite applies.

Insert the Tables 7 and 8 about here

To obtain more detailed information, Table 9 shows the proportion correct score for each item in the group with normal and in the group with aberrant persons as classified by ZU3 (almost the same proportion correct scores were found using C and C^* therefore these proportions are not shown). In particular the arithmetic items 5 through 9 are clearly more difficult for the aberrant examinees than for the normal examinees, whereas the knowledge items 30, 31, 33, and 35 through 37 are clearly easier for the aberrant examinees than for the normal examinees.

Insert Table 9 about here

Finally, the statistic values were related to other experience data. To limit ourselves to ZU3 (however, the same results were obtained using C and C^*), it was found that 69% of the 66 persons with U3>1.30 followed a course in statistics on an easier x-level (as opposed to the more difficult y-level), whereas in the normal group this was only 48%. Personal communication with these students revealed that most of them had been quite unsure about the arithmetic items. Because they were anxious to obtain a sufficient result for the exam, they intensively studied the articles on which the knowledge items were based (assuming that this subject matter could be mastered by studying hard, whereas for answering the arithmetic items successfully you should have a special talent for statistics). As a result of this intensive study, they answered more knowledge items correctly than the average student while as a result of the deficiency of arithmetic skills they answered some of the arithmetic items incorrectly.

Furthermore, preparing themselves for the exam, students had the opportunity to attend extra courses in practising the arithmetic items. From the 21 students that regularly attended these classes there were only two persons that generated an aberrant item score pattern. Thus most persons of this group had no problems with correctly answering the arithmetic items.

Discussion

In this study we compared the power of three person-fit statistics that were defined outside the parametric IRT context. For the simulated data, given a fixed error rate the detection rates of C^* and ZU3 were almost the same, whereas C recovered in general somewhat less simulees. However, for the empirical data the power of the three statistics was approximately the same.

In particular for C, it may be recommended to empirically establish the significance levels for each test. However, empirical establishment of their significance levels requires a large number of examinees. In contrast, ZU3 has theory-based significance levels. Our findings suggest that these levels are in concordance with the critical values found in our simulated data. However, ZU3 is based on the assumption of nondecreasing IRFs. Consequently, it is assumed that the an increase in the number-correct score yields higher probabilities of answering an item correctly. Therefore, items with poor discrimination power should not be selected.

Furthermore, some additional analyses with simulated data showed that if the test mainly consists of items with weak discrimination indices this may inflate the statistic values of both C and C^*. Thus although these indices are

proposed outside the IRT context it is important to select items that allow unidimensional measurement.

It should be noted that we used C and C^* in the context of IRT modeling, although these statistics can also be used outside the IRT context. However, without a theory of measurement it may be difficult to exactly define normal and aberrant persons.

In the empirical example we emphasized the use of person-fit statistics as a diagnostic tool. In our opinion person-fit statistics are a first help to trace persons whose answer behavior is (partly) the result of other characteristics than the particular latent attribute that is assumed to underlie the test results. Other ways of using person-fit analysis are possible. For example, Schmitt, Cortina, and Whitney (1993) used a person-fit statistic to remove persons from the data set that may impair the validity of a test.

In the empirical example we were able to detect persons that performed better on the knowledge items than on the arithmetic items. Persons that perform better on the arithmetic items than on the knowledge items behave as the average student and in general will not be classified as aberrant.

Finally, it should be realized that most person-fit statistics are designed to detect persons that are improbable given an IRT model or given the other persons in the group. If one has reasons to suspect that a particular kind of aberrant behavior underlies the response behavior on the test specialized statistics can be used to detect such persons; for example the statistics discussed by Frary (1993) to detect answer-copying.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, Statistical theories of mental test scores. Reading: Addison-Wesley.

Drasgow F., Levine M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. Applied Psychological Measurement, 11, 59-79.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. British Journal of Mathematical and Statistical Psychology, 38, 67-86.

Frary, R. B. (1993). Statistical detection of multiple-choice answer copying: review and commentary. Applied Measurement in Education, 6, 153-165.

van der Flier, H. (1980). Vergelijkbaarheid van individuele testprestaties [Comparability of individual test performance]. Lisse: Swets & Zeitlinger.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. Journal of Cross-Cultural Psychology, 13, 267-298.

Guttman, L. (1950). The basis for scalogram analysis. In S.A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), Measurement and prediction (pp. 60-90). Princeton: Princeton University Press.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff.

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. Journal of Educational Measurement, 20, 191-205.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 18, 133-146.

Holland, P. W. (1981). When are item response models consistent with observed data? Psychometrika, 46, 79-92.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response theory. Homewood IL: Dow Jones-Irwin.

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. British Journal of Mathematical and Statistical Psychology, 35, 42-56.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. Journal of Educational Statistics, 4, 269-290.

Meijer R. R. (1994). Nonparametric person fit analysis. Unpublished doctoral dissertation. Amsterdam: Vrije Universiteit.

Meijer, R. R., & Sijtsma (in press). Detecting aberrant item score patterns: A review of recent developments. Applied Measurement in Education.

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (in press). Item, test, person and group characteristics and their influence on nonparametric appropriateness measurement. Applied Psychological Measurement.

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. Applied Psychological Measurement, 14, 283-298.

Miller, M. D. (1986). Time allocation and patterns of item response. Journal of Educational Measurement, 23, 147-156.

Mokken, R. J. (1971). A theory and procedure of scale analysis. New York/Berlin: De Gruyter.

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. Applied Psychological Measurement, 6, 417-430.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. Psychometrika, 55, 75-106.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielsen & Lydiche.

Reise, P. R., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. Applied Psychological Measurement, 15, 217-226.

Sato, T. (1975). The construction and interpretation of S-P tables. Tokyo: Meiji Tosho (in Japanese).

Schmitt, N. S., Cortina, J. M, & Whitney, D. J. (1993). Appropriateness fit and criterion-related validity. Applied Psychological Measurement, 17, 143-150.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55, 293-325.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. Psychometrika, 49, 95-110.

Tatsuoka, K. K., & Tatsuoka , M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. Journal of Educational Measurement, 7, 215-231.

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response models. In D.J. Weiss (Ed.), New horizons in testing. New York: Academic Press.

Wright, B. D., & Stone, M. H. (1979). Best test design. Rasch measurement. Chicago: Mesa Press.

Table 1

Critical Values for C, C*, and ZU3

(K=17)

% false aberrants	C	C*	ZU3
1	1.05	.53	2.40
5	.80	.41	1.64
10	.65	.33	1.30

Table 2

Critical Values for C, C*, and ZU3

(K=33)

% false aberrants	C	C*	ZU3
1	.88	.51	2.38
5	.78	.40	1.64
10	.67	.31	1.30

Table 3

Percentage of Guessing Simulees Classified as Aberrant at Three Levels of False Aberrants (K=17)

% false aberrants	C	C*	ZU3
1	38	41	44
5	67	68	70
10	77	80	81

Table 4

Percentage of Cheating Simulees Classified
as Aberrant at Three Levels of False Aberrants
(K=17)

% false aberrants	C	C*	ZU3
1	38	40	43
5	73	76	76
10	84	91	89

Table 5

Percentage of Guessing Simulees Classified as Aberrant at Three Levels of False Aberrants (K=33)

% false aberrants	C	C*	ZU3
1	67	73	72
5	85	89	89
10	90	93	94

Table 6

Percentage of Cheating Simulees Classified
as Aberrant at Three Levels of False Aberrants
(K=33)

% false aberrants	C	C*	ZU3
1	73	79	78
5	90	94	92
10	93	97	98

Table 7

Percentage of Examinees Classified as Aberrant
Using Three Different Cut Scores for C, C^*, and ZU3

C	% examinees	C^*	% examinees	ZU3	% examinees
>.88	4	>.51	3	>2.38	3
>.78	8	>.40	9	>1.64	8
>.67	16	>.33	15	>1.30	15

Table 8

Mean Proportion Correct Answers to the knowledge Items (K) and Arithmetic Items (A) using C, C^* and ZU3

C	K	A	C^*	K	A	ZU3	K	A
$\leq$.67	.49	.63	$\leq$.3	.47	.64	$\leq$ 1.30	.46	.65
>.67	.53	.46	>.3	.54	.47	> 1.30	.54	.47

Table 9

Proportion Correct Score on Each Item in the Group with Normal and in the Group with Aberrant Examinees according to ZU3

	π_g	
item	normal	aberrant
1	.83	.68
2	.83	.60
3	.82	.66
4	.80	.60
5	.79	.36
6	.78	.51
7	.75	.42
8	.75	.40
9	.73	.54
10	.72	.64
11	.72	.63
12	.68	.70
13	.67	.60
14	.66	.65
15	.65	.50
16	.64	.60
17	.60	.65
18	.59	.59
19	.57	.55
20	.55	.60

(continued)

Table 9 (continued)

	π_g	
item	normal	aberrant
21	.51	.53
22	.46	.47
23	.45	.48
24	.41	.44
25	.40	.51
26	.49	.51
27	.43	.44
28	.39	.46
29	.36	.30
30	.36	.67
31	.33	.50
32	.32	.25
33	.31	.50
34	.29	.30
35	.26	.53
36	.23	.54
37	.22	.40

Note: the numbers of the knowledge items are printed bold-faced

Titles of recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-94-12 R.R. Meijer, *Nonparametric and group-based person-fit statistics: A validity study and an empirical example*

RR-94-11 M.P.F. Berger, *Optimal test designs for polytomously scored items*

RR-94-10 W.J. van der Linden & M.A. Zwarts, *Robustness of judgments in evaluation research*

RR-94-9 L.M.W. Akkermans, *Monte Carlo estimation of the conditional Rasch model*

RR-94-8 R.R. Meijer & K. Sijtsma, *Detection of aberrant item score patterns: A review of recent developments*

RR-94-7 W.J. van der Linden & R.M. Luecht, *An optimization model for test assembly to match observed-score distributions*

RR-94-6 W.J.J. Veerkamp & M.P.F. Berger, *Some new item selection criteria for adaptive testing*

RR-94-5 R.R. Meijer, K. Sijtsma & I.W. Molenaar, *Reliability estimation for single dichotomous items*

RR-94-4 M.P.F. Berger & W.J.J. Veerkamp, *A review of selection methods for optimal design*

RR-94-3 W.J. van der Linden, *A conceptual analysis of standard setting in large-scale assessments*

RR-94-2 W.J. van der Linden & H.J. Vos, *A compensatory approach to optimal selection with mastery scores*

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score patterns on the power of a person-fit statistic*

RR-93-1 P. Westers & H. Kelderman, *Generalizations of the Solution-Error Response-Error Model*

RR-91-1 H. Kelderman, *Computing Maximum Likelihood Estimates of Loglinear Models from Marginal Sums with Special Attention to Loglinear Item Response Theory*

RR-90-8 M.P.F. Berger & D.L. Knol, *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models*

RR-90-7 E. Boekkooi-Timminga, *A Method for Designing IRT-based Item Banks*

RR-90-6 J.J. Adema, *The Construction of Weakly Parallel Tests by Mathematical Programming*

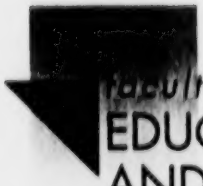
RR-90-5 J.J. Adema, *A Revised Simplex Method for Test Construction Problems*

RR-90-4 J.J. Adema, *Methods and Models for the Construction of Weakly Parallel Tests*

RR-90-2 H. Tobi, *Item Response Theory at subject- and group-level*

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

Research Reports can be obtained at costs from Bibliotheek, Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.



Faculty of
EDUCATIONAL SCIENCE
AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands